Genes involved in human sialic acid biology do not harbor signatures of recent positive selection

Jiyun M. Moon, David M. Aronoff, John A. Capra, Patrick Abbot*, Antonis Rokas*

Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA

*Corresponding Authors: Patrick Abbot; patrick.abbot@vanderbilt.edu / Antonis Rokas; antonis.rokas@vanderbilt.edu

Running title: Recent Selection on Human Sialome Genes

Keywords: Sialic acid, sialome, recent selection, population genetics, human evolution, host immunity

Abstract

Sialic acids are nine carbon sugars ubiquitously found on the surfaces of vertebrate cells and are involved in various immune response-related processes; the overall diversity of sialic acids is often referred to as the host "sialome". In humans, at least 58 genes spanning diverse functions, from biosynthesis and activation to recycling and degradation, are involved in sialic acid biology. Several sialome genes have experienced higher rates of non-synonymous substitutions in the human lineage than their counterparts in other great apes, which may be indicative of ancient positive selection in response to pathogens. To test whether sialome genes have also experienced more recent positive selection in human populations, reflecting adaptation to contemporary cosmopolitan or geographically-restricted pathogens, we calculated several metrics that quantify changes in allele frequency spectra caused by recent selection on the protein-coding and putative enhancer regions of 55 sialome genes using whole genome sequencing data of 2,504 humans from five ethnic groups. To disentangle the effects of demography, we compared the observed patterns in sialome putative enhancer regions and genes to those of 189 housekeeping genes and their putative enhancer regions, which are known to be evolving under strong negative selection due to functional constraints. We found that the patterns of genetic variation of sialome genes and putative enhancers do not differ significantly from those of their housekeeping counterparts. Furthermore, the observed patterns of genetic variation were not significantly different among the four functional categories of sialome genes. These results suggest that the genic and putative enhancer regions of sialome genes have experienced strong purifying selection but not recent positive selection. We propose that the absence of signatures of recent positive selection is consistent with the view that human sialome genes

regulate immune responses against ancient rather than contemporary cosmopolitan or geographically-restricted pathogens.

Introduction

Sialic acids are nine carbon sugars that are commonly found on the ends of glycoconjugates in deuterostomes (Schauer 1982; Varki 2007). These molecules are involved in several biological processes, such as intercellular adhesion and signaling (Varki & Varki 2007; Kelm & Schauer 1997), and play important roles in the modulation of various aspects of the host immune system (Varki & Gagneux 2012; Pilatte et al. 1993) including activation of immune responses, leukocyte trafficking, complement pathway activation, and microbial attachment. Modification of sialic acid molecules during their biosynthesis and subsequent attachments to underlying sugars via different linkages give rise to a diverse repertoire of sialic acids, which is referred to as the "sialome" (Angata & Varki 2002; Varki & Varki 2007; Cohen & Varki 2010; Hayakawa & Varki 2011).

More than 50 genes are known to be involved in various aspects of sialic acid biology in humans, and fall into five broad functional categories (Altheide et al. 2006): 1) biosynthesis, 2) activation, transport, and transfer, 3) modification, 4) recognition, and 5) recycling and degradation (Figure 1). Sialic acid biosynthesis genes are involved in assembling sialic acids from precursor molecules in the cytosol, while genes involved in the second category activate the sialic acid by attaching cytidine monophosphate to them, transport the activated sialic acids to the Golgi for attachment to glycoconjugates via multiple forms of α linkages, and transfer them to the cell surface. Activated sialic acids that are linked to the underlying sugar can be further modified by incorporation of additional molecules before the sialylated glycoconjugate is transferred to the cell surface. At the cell surface, sialic acids are recognized by specialized receptors and finally, the used sialic acids end up in the lysosome for recycling and degradation.

Both the sequence diversity as well as the set of genes involved in human sialic acid biology have been hypothesized to be a result of ancient selective pressures from paleo-pathogens, which have left their signatures on genes involved in sialic acid biology (Varki 2009). For instance, certain strains of Group B *Streptococcus* (GBS) express sialic acids on their cell surfaces and can bind to inhibitory receptors such as Siglec-5, leading to subsequent down regulation of host immune responses (Carlin et al. 2009). It is thought that the binding of GBS to Siglec-5 in an ancestor of Old World monkeys drove the evolution of Siglec-14, a receptor with identical recognition capacity as Siglec-5, but with immunostimulatory rather than immunosuppressive activity (Angata et al. 2006; Ali et al. 2014). Another example involves the Alu-mediated inactivation of the *CMAH* gene specifically in the human lineage (Varki 2001; Chou et al. 1998; Irie et al. 1998): the enzyme encoded by this gene is responsible for the conversion of N-acetylneuraminic acid (Neu5Ac) to N-glycolylneuraminic acid (Neu5Gc), and Neu5Ac is the main form of sialic acid in humans. This change in sialic acid composition has been suggested to act as a means to escape infection by *Plasmodium reichenowi*, a parasite that preferentially binds to Neu5Gc and is known to cause malaria in chimpanzees. Conversely, adoption of Neu5Ac likely made humans susceptible to the malaria parasite *Plasmodium falciparum* infection, which instead binds to Neu5Ac (Martin et al. 2005; Varki & Gagneux 2009). Finally, molecular evolutionary analyses in primates and rodents have shown that genes involved in sialic acid recognition show a considerable degree of sequence divergence, even between closely related species, suggestive of ancient positive selection in response to pathogens (Altheide et al. 2006).

The examples above suggest that sialome genes experienced adaptive changes early in the radiation of primates (including hominids). However, they do not address whether these same genes have also experienced recent positive selection in human populations, reflective of adaptation to contemporary pathogens. To test this hypothesis, we estimated the population-level genetic variation of both the protein-coding and putative enhancers of sialome genes, and carried out tests of neutrality using the 1000 Genomes Project sequencing data from 2,504 humans belonging to five ethnic groups. To characterize the relative strength of signatures of selection on sialome genes and to account for the potentially cofounding effect of past demographic events, we further compared the signatures of recent positive selection of 55 sialome genes to 189 housekeeping genes. Finally, we compared the patterns of genetic variation among functional categories of sialome genes. Our results reveal the presence of purifying selection acting ubiquitously across the sialome, and the absence of recent positive selection. Both protein-coding and putative enhancer regions of sialome genes, irrespective of their functional category, have similar patterns of genetic variation as housekeeping human genes. Given the evidence of ancient selection on sialome genes prior to the emergence of modern humans (Altheide et al. 2006), we hypothesize the evolution of human sialome genes has been more strongly influenced by ancient primate pathogens rather than by contemporary cosmopolitan or geographically-restricted human pathogens.

Methods

Genotype Dataset

To examine whether sialome genes have experienced recent positive selection in human populations, we used the genotype dataset of Phase 3 of the 1000 Genomes Project (Auton et al.

2015), which included whole genome sequence data for 2,504 individuals from five ethnic backgrounds (Africans, Europeans, East Asians, South Asians, and admixed Americans) for all analyses (for information on data source, see File S1).

Sialome Genes

To identify all the genes involved in sialic acid biology, we started from a previously published list of 55 loci (Altheide et al. 2006) and added *SIGLEC14* and *SIGLEC16*, which were discovered more recently, as well as *SIGLEC15*, which was not included in the previous study. As sex-linked genes tend to exhibit different patterns of genetic variation than genes located on autosomal chromosomes (Schaffner 2004), we excluded two genes (*L1CAM* and *RENBP*) that are located on the X chromosome; in addition, we removed *CMAH*, the sole gene involved in modification of newly synthesized sialic acids, because it is a non-functional pseudogene in humans (Chou et al. 1998; Irie et al. 1998). Thus, 55 genes were retained for subsequent analyses (Figure 1).

To extract genotypes for the loci of interest, we used *VCFtools*, version 0.1.13 (Danecek et al. 2011) (for specific commands used, see File S1), with the gene coordinates (GRCh37p.13) obtained from Ensembl (release 75) via BioMart as input; indels were not included in our analyses. Compressed VCF files and tab-delimited index files required for subsequent analyses were created using *tabix*, version 0.2.6 (Li 2011).

Putative Enhancers Associated with Sialome Genes

To identify putative enhancers that are adjacent to sialome genes, we used data on predicted enhancers associated with putative target genes from the FANTOM5 Project (Andersson et al. 2014). To intersect sialome gene coordinates with putative target transcription start sites we used *BEDtools*, version 2.26.0 (Quinlan & Hall 2010). To extract genotype corresponding to these putative enhancer regions, we used *VCFtools*, as described above (for information on specific commands used, see File S1). Four genes, namely *CMAS*, *LAMA1*, *NANP*, and *ST8SIA1*, were excluded from the analyses, as there were no genotype data associated with their putative enhancer regions in the 1000 Genomes Project dataset.

Housekeeping Genes

Population demographic events can mimic the effect of genuine positive selection on allele frequency spectra (Sabeti et al. 2006). To disentangle the confounding effects of demography, we also examined the patterns of genetic variation in housekeeping genes and compared them to those of the sialome genes. We obtained all genes belonging to one of the following three functional protein classes: 1) RNA polymerase related proteins, 2) ribosomal proteins, and 3) citric acid cycle related proteins, and that were expressed in all tissues (n = 44; Transcripts Per Kilobase Million (TPM) ≥ 1) from the Human Protein Atlas Database, version 16.1 (Uhlén et al. 2015). The 195 genes (RNA polymerase related proteins: n = 25; ribosomal proteins: n = 144; citric acid cycle related proteins: n = 26) included in this initial list were further filtered to include only genes on autosomal chromosomes (n = 189; RNA polymerase related proteins: n = 25; ribosomal proteins: n = 140; citric acid cycle related proteins: n = 24). Gene coordinates of housekeeping genes and associated putative enhancer regions were obtained as described above. Twenty-three housekeeping genes were excluded from subsequent analyses

of enhancer regions due to either missing enhancer data in the FANTOM5 project dataset or lack of genotype data within the enhancer regions in the 1000 Genomes Project dataset.

Estimation of genetic variation and detection of signatures of recent positive selection

To determine the levels of genetic polymorphism of sialome genes, we first calculated pairwise nucleotide diversity (π) (Nei & Li 1979) using *PopGenome*, version 2.1.6 (Pfeifer et al. 2014). As there may have been episodic variation in the occurrence and duration of recent positive selection, we next used several different tests aimed to capture any signatures left from the action of selection at different evolutionary time scales (Sabeti et al. 2006). Specifically, to detect signatures left by recent selective events that occurred ~250,000 years ago, we calculated Tajima's *D* (Tajima 1989) and Fu & Li's *D** and *F** values (Fu & Li 1993; Fu 1997) using the R package *PopGenome*. To assess the statistical significance of these neutrality test indices, we compared the observed distribution of values against a distribution of values "expected" under assumptions of neutrality. This "expected" distribution of values was obtained by first simulating 1,000 sequences with the mutation parameter (Θ or Tajima's theta) fixed using the *ms* software (Hudson 2002); (for specific commands used, see File S1). We next used these simulated sequences to calculate Tajima's *D*, Fu & Li's *D** and *F** values using *PopGenome*. We defined a *p*-value for each test index as the proportion of values on simulated sequences that were more extreme (i.e. negative) than the observed value. As significant deviations from neutrality could be due to selective sweeps associated with positive selection, background selection or population expansion, we also calculated Fay & Wu's *H* values (Fay & Wu 2000) to determine if significantly negative values for the above three indices were caused by genuine positive selection using *PopGenome* (for specific commands used, see File S1).

Finally, to detect recent local selective events that occurred after the out-of-African migration event approximately 75,000 to 50,000 years ago, we calculated both weighted and mean Weir & Cockerham's F_{ST} (Weir & Cockerham 1984) at both the global (comparisons of all five ethnic groups) and pairwise (comparisons of one ethnic group against the other four) levels using *VCFtools* (Danecek et al. 2011).

Statistical Analysis

To compare the patterns of nucleotide diversities (one-tailed), the three neutrality test indices (one-tailed), and F_{ST} (one-tailed) between the sialome genes and the housekeeping genes, as well as to compare the patterns of genetic variation across different functional categories of sialome genes (two-sided), we conducted Mann-Whitney *U* (MW*U*) tests: the exact *p*-values were calculated for each comparison via the *coin* package, version 1.1-3, in the R programming environment (Hothorn et al. 2008). *P*-values calculated for comparisons among functional categories were adjusted *post hoc* to correct for the testing of multiple hypotheses via the Bonferroni method using R. To determine the statistical significance of the enrichment of a particular functional category in the top 100 SNPs (i.e., the 100 SNPs showing the highest F_{ST} values), we carried out hypergeometric tests using the *phyper* function of R.

Results

Sialome genes do not exhibit significantly stronger signatures of recent positive selection than housekeeping genes

Comparison of pairwise π between sialome and housekeeping genes showed that, in general, the values of sialome genes were significantly higher than those of housekeeping genes (Figure 2A & Figure S1; All individuals: one-sided MW*U* = 6,706, *p*-value < 0.001; Africans: one-sided MW*U* = 6,580, p-value = 0.001; Europeans: one-sided MW*U* = 6,521, *p*-value = 0.002; East Asians: one-sided MW*U* = 7,154, *p*-value < 0.001; South Asians: one-sided MW*U* = 6,702, *p*-value < 0.001; Americans: one-sided MW*U* = 6,736, *p*-value < 0.001). However, only 3-4 sialome genes displayed π values higher than the 95th percentile of the housekeeping genes (Figure 2A & Figure S1), with *SIGLEC12* and *SIGLEC16* exhibiting the highest values in all ethnic groups, *NEU4* in all except Africans, and *SLC17A5* in analyses of all individuals, Africans, and Americans (Table S1).

To test the hypothesis that sialome genes experienced recent positive selection in human populations approximately 250,000 years ago, we calculated Tajima's *D* and Fu & Li's *D** and *F** over all 2,504 individuals (Figure 2B &Table S2). All sialome genes exhibited a statistically significant excess of singletons as shown by the significantly negative values of both Fu & Li's *D** or *F**; similarly, all genes but *SLC75A1* and *SIGLEC14* displayed significantly negative values of Tajima's *D*. However, comparison of the values for these three indices between sialome and housekeeping genes showed that the distributions of values for the two sets of genes were somewhat similar: with the exception of Fu & Li's *D** (MW*U* = 3170, *p*-value < 0.001), there were no significant differences between the values of the neutrality indices of the sialome and housekeeping genes. In addition, only one (*ST3GAL2*), four (*SIAE*, *ST3GAL2*, *ST8SIA4*, *LAMA2*) and one (*SIGLEC8*) genes exhibited lower values compared to the housekeeping genes for Tajima's *D* and Fu & Li's *D** & *F**, respectively. To examine whether any of these 6 genes

show an excess of high frequency derived variants, we calculated Wu & Fay's *H* (Fay & Wu 2000). None of the 55 sialome genes displayed negative Fay & Wu's *H* values, ruling out positive selection as a possible mode of evolution around 250,000 years ago for the sialome genes.

To detect signatures of local positive selection that occurred after the major human migration out of Africa, we calculated Weir & Cockerham's F_{ST} for the sialome and housekeeping genes. We found that most sialome genes do not exhibit significantly higher levels of population differentiation than housekeeping genes (Figure 2C & Figure S2). A considerable number of genes exhibited F_{ST} values lower than the median of the housekeeping genes. For example, 34/55 (61.82%) and 32/55 (58.18%) genes displayed lower global and pairwise F_{ST} values in the African population than the median of the housekeeping genes, respectively. Similar results were obtained for the other human populations, with 36-50% of pairwise F_{ST} values of the sialome genes calculated for the respective ethnic groups exhibiting values lower than the median of the housekeeping genes. Only eight genes exhibited F_{ST} values higher than the 95th percentile of housekeeping genes in each population examined; *NANP* in all 2,504 individuals, *SIAE* in Africans, *CMAS* in Europeans, *NANP* in East Asians, *CD33* in Americans, and *NANS*, *SIGLEC11*, and *NEU3* in South Asians (Table S1).

As expression levels of genes involved in immune response vary within and between populations (Nédélec et al. 2016), we also tested the hypothesis that enhancer regions adjacent to sialome genes have experienced recent positive selection by calculating the same statistics that we used to analyze the genic regions. As control, we used predicted enhancer regions in

proximity to the housekeeping genes. We found no statistically significant differences between the nucleotide diversity values of putative enhancers of 51 sialome genes and those of the 166 housekeeping genes, with none of the putative sialome enhancers displaying values that exceeded the 95th percentile of the housekeeping ones (Figure 3A, Figure S3 & Table S3).

Calculation of all three neutrality indices showed that most putative sialome enhancers were associated with significantly negative values (Figure 3B & Table S4). Specifically, putative enhancers adjacent to only five (*NAGK*, *NEU4*, *NPL*, *SIGLEC1*, and *ST3GAL6*) and two (*NEU2* and *ST3GAL4*) genes displayed Tajima's *D* and Fu & Li's *D** values that exhibited non-significant deviation from neutrality, respectively: putative enhancers of all 51 genes exhibited significantly negative values of Fu & Li's *F** values. While the values of Fu & Li's *D** and *F** of the predicted sialome enhancers overall were significantly lower than those of the housekeeping counterparts (Fu & Li's *D**: MW*U* = 3501, *p*-value = 0.03507; Fu & Li's *F**: MW*U* = 1066, *p*-value < 0.001), most sialome enhancers did not exhibit "extremely negative" values compared to the 5th percentile values of the three neutrality indices of putative enhancers adjacent to housekeeping genes. Only one putative sialome enhancer (*NEU3*) had a lower Tajima's *D* value compared to the 5th percentile value of housekeeping enhancers; in addition, 12 and 11 putative sialome enhancers displayed Fu & Li's *D** and *F** values lower than the 5th percentile values of housekeeping enhancers, respectively.

We finally tested the hypothesis that enhancers adjacent to sialome genes have experienced positive selection that occurred after the major human migration out of Africa. We found no significant differences between the F_{ST} values of putative sialome and housekeeping

enhancers. Furthermore, none of the F_{ST} values of the putative sialome enhancers exceeded the 95th percentile values of the putative housekeeping enhancers (Figure 3C & Figure S4), with the exceptions of *ST3GAL5* in Europeans, *NEU2*, *ST6GALNAC4* and *ST6GALNAC6* in East Asians, and *CTSA* in Americans (Table S3).

The four functional categories of sialome genes do not significantly differ in their signatures of recent positive selection

We next examined whether four distinct functional categories of sialome genes exhibit statistically significant differences in their allele frequency spectra by comparing the patterns of nucleotide diversity, neutrality test indices and F_{ST} values across them. Comparison of π values did not reveal significant differences among the four functional categories (Figure 4A & Table S5). Similar patterns were observed when we compared the values of the three neutrality indices; the only exception was the comparison of Fu & Li's *D** values between the activation, transport, transfer and recognition categories (MW*U* = 100, adjusted *p*-value = 0.021) (Figure 4B & Table S5). Similarly, examination of F_{ST} values showed that patterns of population differentiation were similar for the four functional categories of sialome genes tested (Figure 4C & Table S5).

We next examined if putative enhancers associated with genes belonging to the four functional categories of sialic acid biology differed significantly in their allele frequency spectra. There were statistically significant differences in π values only between recognition and recycling, degradation categories (All individuals: MW*U* = 115, adjusted *p*-value = 0.02; Africans: MW*U* = 117, adjusted *p*-value = 0.012; Europeans: MW*U* = 113, adjusted *p*-value = 0.032; East Asians: MW*U* = 111, adjusted *p*-value = 0.048) and between activation, transport,

transfer and recognition categories (All individuals: MW*U* = 90, adjusted *p*-value = 0.044; Europeans: MW*U* = 83, adjusted *p*-value = 0.022; South Asians: MW*U* = 88, adjusted *p*-value = 0.036; Americans: MW*U* = 80, adjusted *p*-value = 0.016) (Figure 5A & Table S6). Furthermore, no significant differences were observed in comparisons of the values of the three neutrality indices (Figure 5B & Table S6) or of F_{ST} values (Figure 5C &Table S6) of putative enhancers across the four functional categories, with the exception of F_{ST} values calculated for the biosynthesis and recognition categories in all 2,504 individuals (MW*U* = 38, adjusted *p*-value = 0.029).

Sialome SNPs with the highest degree of population differentiation

If selective pressures occurred only within specific regions within a gene, estimates of genetic variation based on entire genic regions could dilute signatures of recent positive selection. For example, it has been previously shown that the first Ig-like domains (Ig1, V set Ig-like domain) of sialic acid binding lectins (*SIGLEC*s) are responsible for the high average Ka/Ks value of genes belonging to the recognition category (Altheide et al. 2006). Therefore, we also examined single SNP estimates of Weir & Cockerham's F_{ST} and ranked the SNPs that reside within the genic regions of the sialome genes according to their F_{ST} values. When we examined all 2,504 individuals, most of the top 100 SNPs belonged to recognition genes (56%), followed by activation, transport, transfer genes (39%), and recycling, degradation genes (5%) (Table 1). None of the SNPs of the biosynthesis genes were in the top 100. A SNP within the genic region of *LAMA1* exhibited the highest F_{ST} value (0.613). Examination of single SNP estimates of pairwise F_{ST} values calculated for each ethnic group showed similar patterns: the most common

SNPs of genes in the top 100 were involved in activation, transport, transfer and recognition (Table S7).

As the activation, transport, transfer and recognition categories have the most genes (n = 22 and n = 21, respectively) and the most genic SNPs (69,807 and 35,945 out of 114,832 SNPs, respectively), this apparent enrichment could simply be an artifact due to the high number of SNPs within these two functional categories. Therefore, we carried out hypergeometric tests to determine if the enrichment of the top 100 hits with SNPs of these two functional categories is statistically significant. Enrichment of genic SNPs of the activation, transport, transfer category was statistically significant only in Europeans (p-value = 0.013), and enrichment of genic SNPs in the recognition only in Africans (p-value < 0.001) and in all 2,504 individuals (p-value < 0.001).

We next studied F_{ST} values of SNPs of the genes in the activation, transport, transfer and recognition categories separately. In the activation, transport, transfer category, *ST8SIA1* exhibited the highest single SNP estimate of F_{ST} in Africans, East Asians, and South Asians (Africans: 0.728; East Asians: 0.671; South Asians: 0.395), while *ST6GAL2* (0.294) and *ST3GAL4* (0.405) displayed the highest SNP estimate of F_{ST} in Europeans and Americans, respectively. Finally, *ST6GALNAC3* (0.573) exhibited the highest SNP estimate of global F_{ST}. Among the genes in the recognition category, *LAMA1* exhibited the highest SNP estimate of F_{ST} except in East Asians and South Asians (All individuals: 0.613; Africans: 0.750; Europeans: 0.300; Americans: 0.336). Interestingly, *SIGLEC14*, which is known as the paired immune-stimulatory receptor of *SIGLEC5* exhibited the highest SNP estimates of F_{ST} in East Asians (F_{ST}

= 0.625); this is likely associated with the high frequency of the null deletion polymorphism of the activating receptor *SIGLEC14* in individuals of this ethnic background (Angata et al. 2006). In South Asians, a SNP associated with *CD33* exhibited the highest single SNP estimate of F_{ST} of the recognition genes (0.228).

Finally, we studied the functions of the SNPs with the top 100 F_{ST} estimates using the functional annotation dataset curated by the 1000 Genomes Project Consortium using Ensembl's Variant Effect Predictor (based on release 75). In all ethnic groups studied, 80-90% of the SNPs were intronic variants (All individuals: 93%; Africans: 95%; Europeans: 84%; East Asians: 86%; South Asians: 90%; Americans: 80%). Only 1-2 SNPs were annotated as missense variants and except for one gene (*NEU2* in Europeans), the genes associated with these SNPs were receptors for sialic acids (All individuals: *SELP* (rs6133); Africans: *SELP* (rs6133), *LAMA1* (rs543355); Europeans: *CD33* (rs35112940); East Asians: *SIGLEC12* (rs6509544); South Asians: *SELP* (rs6133), *SIGLEC11* (rs45438992); Americans: *CD33* (rs12459419), *SIGLEC6* (rs35931837)). Only one SNP within the coding region of CD33 (rs12459419: F_{ST}: 0.245686) was annotated as 'deleterious' by SIFT (McLaren et al. 2016).

Discussion

In this study, we tested the hypothesis that loci involved in human sialic acid biology have experienced recent positive selection. To test this hypothesis, we employed several population genetic tests designed to detect signatures of positive selection that occurred at different time points after the origin of modern humans approximately 250,000 years ago. We further compared the calculated metrics to those of housekeeping genes, which are known to

have been evolving under purifying selection (Zhang & Li 2003). Overall, we rarely observed sialic acid genes outside the range of values observed in housekeeping genes. Similarly, we did not observe that the protein-coding and putative enhancer regions of loci belonging to the four functional categories of sialic acid biology have been subject to different selective pressures. Finally, examination of the 100 SNPs with the highest degree of population differentiation showed that most were intronic variants associated with genes in the recognition and activation, transport, transfer functional categories.

The most likely explanation of our results is that the sialome genes have not been targets of positive selection in the last 250,000 years of human evolution. Previous studies have shown that sialome genes were subject to strong ancient selective pressures that have resulted in species-specific adaptations (Varki 2001; Angata et al. 2004; Altheide et al. 2006; Varki & Gagneux 2009), suggesting that they likely evolved in response to ancient pathogens The function of sialome genes, i.e., the generation of host-specific sialic acids and recognition of those self-associated molecular patterns (SAMPs) by cognate receptors, contributes to the detection of missing self (Medzhitov & Janeway 2002). It is possible that evolution of such markers of normal self, and the ability to recognize them correctly, have been influenced by ancient, but not, recent pathogens. For instance, the human-specific Alu-mediated inactivation of *CMAH*, the protein product of which is responsible for generating Neu5Gc from its precursor Neu5Ac, has been suggested as a means of escaping infection by *Plasmodium reichenowi*, a malaria-causing parasite that infects other great apes (Martin et al. 2005; Varki & Gagneux 2009). In addition, human-specific mutation of an essential arginine residue required for sialic acid binding in Siglec-12 has been suggested to be evolutionarily related to the loss of Neu5Gc

in humans (Varki 2009). Several other human-specific changes in sialic acid biology (e.g., deletion of Siglec-13, increased expression of Siglec-1 on human macrophages compared to other primates) could have resulted from selective pressures exerted by pathogens in more ancient time scales (Varki 2009; Wang et al. 2012). It should be noted that several other innate-immunity genes have been shown to lack any signatures of positive selection in more recent evolutionary time scales, including the Toll-Like Receptors (*TLR*s), *MYD88*, and *TRIF* (Mukherjee et al. 2009; Siddle & Quintana-Murci 2014).

Consistent with the interpretation that sialome genes have largely experienced purifying selection during recent human evolution is that they did not originally evolve to participate in the host-pathogen interaction. Host-pathogen interactions can originate via two evolutionary routes (Medzhitov & Janeway 2002). The first route includes cases in which host pattern recognition receptors evolve to recognize molecular patterns of pathogens that are essential to the pathogens' physiology; in such cases, the host receptors will evolve rapidly to counteract any changes occurring in the pathogens' molecular patterns, whose own change is constrained by their involvement in the pathogen's physiology. The second route includes cases in which pathogen genes evolve to hijack host processes that serve functions unrelated to host-pathogen interactions; in this case, it is likely that the pathogen hijacking genes will be the ones evolving rapidly and the genes associated with the hijacked host process will be those whose evolution is constrained (due to their involvement in host biology unrelated to pathogen defense). The host-pathogen interaction between host sialome genes and their pathogens evolved from this second route. Sialome genes are involved in other various biological processes such as intercellular signaling,

development of the nervous system (Schnaar et al. 2014), and establishment of the baseline level of immunity (Chang & Nizet 2014) that are most likely under the influence of purifying selection.

An alternative explanation that would be consistent with our results is that sialome genes were targets of recent positive selection, but there has not been enough time for the new advantageous variants to increase in frequency in the human population; especially so if the selective pressures exerted by pathogens have not been sufficiently strong to result in substantial changes in allele frequency spectra since the emergence of modern humans (Pritchard et al. 2010). Furthermore, methods relying on changes in the allele frequency distributions among linked, neutral polymorphisms have little power to detect soft selective sweeps in which standing variation or multiple *de novo* mutations introduce several beneficial alleles into a population (Pennings & Hermisson 2006a; Messer & Petrov 2013). Such soft selective sweeps leave different signatures on linked variation than hard selective sweeps and are more challenging to detect. Specifically, while classical hard selective sweeps result in reduction of genetic diversity around the selected site and excess of rare singletons, soft selective sweeps do not necessarily produce these signatures; instead, soft selective sweeps can result in presence of more than one, independent ancestral haplotype in the population due to multiple beneficial variants bringing along with them different haplotypes (Pennings & Hermisson 2006b). Alternatively, sialome genes could have been evolving as a result of polygenic adaptation, in which rapid adaptation is possible via modest allele frequency changes in multiple, related loci (Pritchard et al. 2010). For example, it is conceivable that modest positive selection events occurring on numerous sialome SNPs or genes contribute to the host's response to contemporary pathogens; such selection could have stemmed from either soft selective sweeps or polygenic adaptation, resulting in only modest

allele frequency shifts in individual sialome loci, which would be undetectable by the methods used here.

Acknowledgements

We thank Julie Phillips for helpful discussions on the evolution of genes involved in sialic acid biology. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. MM was supported by the Graduate Program in Biological Sciences at Vanderbilt University. This work was supported in part by the March of Dimes through the March of Dimes Prematurity Research Center Ohio Collaborative and by the National Science Foundation (DEB-1442113 to A.R.).

References

Ali SR et al. 2014. Siglec-5 and Siglec-14 are polymorphic paired receptors that modulate neutrophil and amnion signaling responses to group B Streptococcus. J Exp Med. 211:1231–1242.

Altheide TK et al. 2006. System-wide Genomic and Biochemical Comparisons of Sialic Acid Biology Among Primates and Rodents: EVIDENCE FOR TWO MODES OF RAPID EVOLUTION. J Biol Chem. 281:25689–25702.

Andersson R et al. 2014. An atlas of active enhancers across human cell types and tissues. Nature. 507:455–461.

Angata T, Hayakawa T, Yamanaka M, Varki A, Nakamura M. 2006. Discovery of Siglec-14, a novel sialic acid receptor undergoing concerted evolution with Siglec-5 in primates. FASEB J. 20:1964–1973.

Angata T, Margulies EH, Green ED, Varki A. 2004. Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. PNAS. 101:13251–13256.

Angata T, Varki A. 2002. Chemical Diversity in the Sialic Acids and Related α-Keto Acids: An Evolutionary Perspective. Chem. Rev. 102:439–470.

Auton A et al. 2015. A global reference for human genetic variation. Nature. 526:68–74.

Carlin AF et al. 2009. Group B Streptococcus suppression of phagocyte functions by protein-mediated engagement of human Siglec-5. J Exp Med. 206:1691–1699.

Chang YC, Nizet V. 2014. The interplay between Siglecs and sialylated pathogens. Glycobiology. 24:818–825.

Chou HH et al. 1998. A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. Proc. Natl. Acad. Sci. USA. 95:11751–11756.

Cohen M, Varki A. 2010. The Sialome—Far More Than the Sum of Its Parts. OMICS. 14:455–464.

Danecek P et al. 2011. The variant call format and VCFtools. Bioinformatics. 27:2156–2158.

Fay JC, Wu CI. 2000. Hitchhiking Under Positive Darwinian selection. Genetics. 155:1405–1413.

Fu YX. 1997. Statistical tests of neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. Genetics. 147:915-925.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics. 133:693-709.

Hayakawa T, Varki A. 2011. Human-Specific Changes in Sialic Acid Biology in Primatology Monographs, pp.123-148 in *Post-Genome Biology of Primates*, edited by H. Hirai et al. Springer.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A. 2008. Implementing a Class of Permutation Tests: The coin Package. Journal of Statistical Software; 28:1-23.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18:337–338.

Irie A, Koyama S, Kozutsumi Y, Kawasaki T, Suzuki A. 1998. The Molecular Basis for the Absence of N-Glycolylneuraminic Acid in Humans. J Biol Chem. 273:15866–15871.

Kelm S, Schauer R. 1997. Sialic Acids in Molecular and Cellular Interactions. Int Rev of Cytol 175:137–240.

Li H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 27:718–719.

Martin MJ, Rayner JC, Gagneux P, Barnwell JW, Varki A. 2005. Evolution of human-chimpanzee differences in malaria susceptibility: Relationship to human genetic loss of *N*-glycolylneuraminic acid. Proc. Natl. Acad. Sci. USA. 102:12819–12824.

McLaren W et al. 2016. The Ensembl Variant Effect Predictor. Genome Biology. 17:122.

Medzhitov R, Janeway CA. 2002. Decoding the Patterns of Self and Nonself by the Innate Immune System. Science. 296:298–300.

Messer PW, Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends in Ecology & Evolution. 28:1–11.

Mukherjee S, Sarkar-Roy N, Wagener DK, Majumder PP. 2009. Signatures of natural selection are not uniform across genes of innate immune system, but purifying selection is the dominant signature. Proc. Natl. Acad. Sci. USA. 106:7073–7078.

Nei M. Li, WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA. 76:5269–5273.

Nédélec Y et al. 2016. Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell. 167:657–669.e21

Pennings PS, Hermisson J. 2006a. Soft Sweeps II--Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. Mol Biol Evol. 23:1076–1084.

Pennings PS, Hermisson J. 2006b. Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation. PLoS Genet. 2:e186.

Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. Mol Biol Evol. 31:1929–1936.

Pilatte Y, Bignon J, Lambré CR. 1993. Sialic acids as important molecules in the regulation of the immune system: pathophysiological implications of sialidases in immunity. Glycobiology. 3:201–218.

Pritchard JK, Pickrell JK, Coop G. 2010. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Current Biology. 20:R208–R215.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842.

Sabeti PC et al. 2006. Positive Natural Selection in the Human Lineage. Science. 312:1614–1620.

Schaffner SF. 2004. The X chromosome in population genetics. Nature Rev Genet. 5:43–51.

Schauer R. 1982. Chemistry, Metabolism, and Biological Functions of Sialic Acids. Adv. Carbohydr. Chem. Biochem. 40:31–234.

Schnaar RL, Gerardy-Schahn R, Hildebrandt H. 2014. Sialic Acids in the Brain: Gangliosides and Polysialic Acid in Nervous System Development, Stability, Disease, and Regeneration. Physiol Rev. 94:461–518.

Siddle KJ, Quintana-Murci L. 2014. The Red Queen's long race: human adaptation to pathogen pressure. Curr. Opin. Genet. Dev. 29:31–38.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585-595.

Uhlén M et al. 2015. Tissue-based map of the human proteome. Science. 347:1260419–1260419.

Varki A. 2007. Glycan-based interactions involving vertebrate sialic-acid-recognizing proteins. Nature. 446:1023–1029.

Varki A. 2001. Loss of N-glycolylneuraminic acid in humans: Mechanisms, consequences, and implications for hominid evolution. Am. J. Phys. Anthropol. 116:54–69.

Varki A. 2009. Multiple changes in sialic acid biology during human evolution. Glycoconj J. 26:231–245.

Varki A, Gagneux P. 2009. Human-specific evolution of sialic acid targets: Explaining the malignant malaria mystery? Proc. Natl. Acad. Sci. USA. 106:14739–14740.

Varki A, Gagneux P. 2012. Multifarious roles of sialic acids in immunity. Ann. N.Y. Aca. Sci.

1253:16–36.

Varki NM, Varki A. 2007. Diversity in cell surface sialic acid presentations: implications for biology and disease. Lab Invest. 87:851–857.

Wang X et al. 2012. Specific inactivation of two immunomodulatory *SIGLEC* genes during human evolution. Proc. Natl. Acad. Sci. USA. 109:9935–9940.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evol. 38:1358–1370.

Zhang L, Li WH. 2003. Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes. Mol Biol Evol. 21:236–239.

Figure legends

Figure 1. Summary of the biochemical processes involved in human sialic acid biology and the genes known to be associated with them

Figure 2. Sialome genes do not exhibit significantly stronger signatures of recent positive selection compared to housekeeping genes

A. The graph presents the nucleotide diversity (π) values calculated over all 2,504 individuals of the 1000 Genomes Project. The gray step-histogram depicts the distribution of the π values of the 189 housekeeping genes. The red dot-dashed line indicates the 95th percentile π value of the housekeeping genes. The vertical blue lines correspond to the individual π values of the 55 sialome genes. The values of the nucleotide diversity for each gene can be found in Table S1

B. The graph presents the values of the three neutrality indices calculated for all 2,504 individuals of the 1000 Genomes Project. For each neutrality test, the gray step-histogram depicts the distribution of the values of the 189 housekeeping genes. The red dot-dashed line indicates the 5th percentile value of the housekeeping genes. The vertical blue lines correspond to the individual values of the three neutrality indices for the 55 sialome genes. The values of the three neutrality indices for each gene can be found in Table S2.

C. The graph presents weighted Weir & Cockerham's F_{ST} values calculated across all five ethic groups of the 1000 Genomes Project. The gray step-histogram depicts the distribution of the F_{ST} values of the 189 housekeeping genes. The red dot-dashed line indicates the 95th percentile F_{ST} value of the housekeeping genes. The vertical blue lines correspond to the individual F_{ST} values

of the 55 sialome genes. The values of the F_{ST} for each gene, along with the values of the mean Weir & Cockerham's F_{ST}, can be found in Table S1.

Figure 3. Putative enhancers of sialome genes do not exhibit significantly stronger signatures of recent positive selection compared to housekeeping genes

A. The graph presents the nucleotide diversity (π) values calculated for all 2,504 individuals of the 1000 Genomes Project. The gray step-histogram depicts the distribution of the π values of the putative enhancers of the 166 housekeeping genes. The red dot-dashed line indicates the 95th percentile π value of the putative housekeeping enhancers. The vertical blue lines correspond to the individual π values of the 51 putative sialome enhancers. The π values for each putative enhancer can be found in Table S3.

B. The graph presents the values of the three neutrality indices calculated for all 2,504 individuals of the 1000 Genomes Project. For each neutrality test, the gray step-histogram depicts the distribution of the values of the putative enhancers of the 166 housekeeping genes. The red dot-dashed line indicates the 5th percentile value of the putative housekeeping enhancers. The vertical blue lines correspond to the individual values of the three neutrality indices for the putative enhancers of the 51 sialome genes. The values of the three neutrality indices of individual putative enhancers can be found in Table S4.

C. The graph presents weighted Weir & Cockerham's F_{ST} values calculated across all five ethic groups of the 1000 Genomes Project. The gray step-histogram depicts the distribution of the F_{ST} values of the putative enhancers of the 166 housekeeping genes. The red dot-dashed line

indicates the 95th percentile F_{ST} value of the putative housekeeping enhancers. The vertical blue lines correspond to the individual F_{ST} values of the 51 putative sialome enhancers. The F_{ST} values for each putative enhancer, along with the values of the mean Weir & Cockerham's F_{ST}, can be found in Tables S3.

Figure 4. Genes in the four functional categories of sialic acid biology do not significantly differ in their signatures of recent positive selection

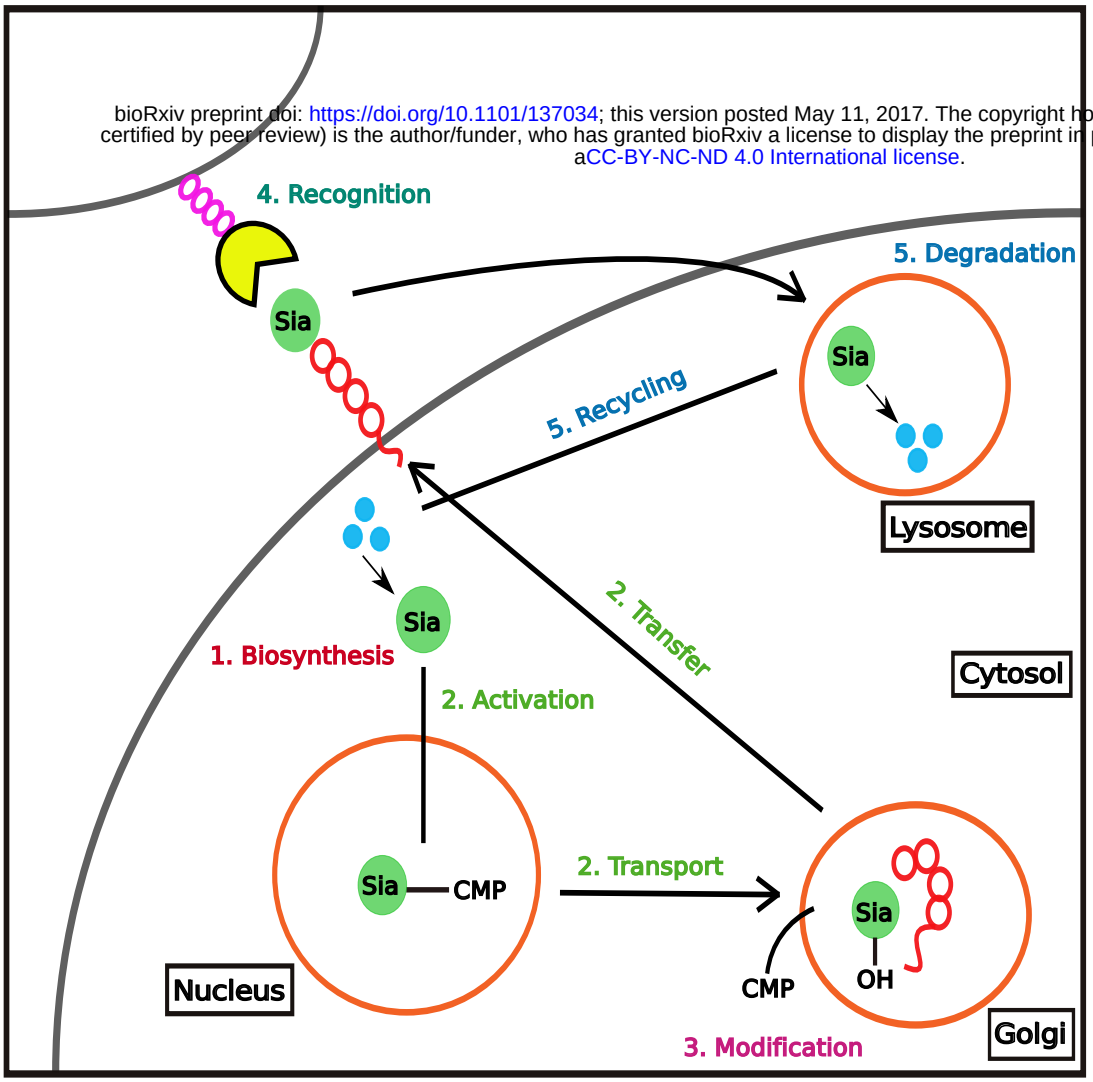
The averages of the values for nucleotide diversity (A), the three neutrality indices (B), and Weir & Cockerham's F_{ST} (C) for each functional category are shown. Each symbol represents the average value for each population group. For nucleotide diversity (A) and Weir & Cockerham's F_{ST} values (C), solid gray lines indicate the median values of the 189 housekeeping genes calculated for all 2,504 individuals; dash-dotted gray lines represent the median values of the 189 housekeeping genes calculated for Africans; the gray dotted lines, loosely-dashed lines, densely-dashed, and dashed lines indicate the median values of the 189 housekeeping genes calculated for Europeans, East Asians, South Asians, and Americans, respectively. For the neutrality indices (B), the gray dash-dotted line represents the median values of the 189 housekeeping genes calculated for all 2,504 individuals. The number of genes belonging to each category is shown below the name of each functional category.

Figure 5. There are no significant differences in signatures of recent positive selection among the putative enhancers of the four functional categories of sialic acid biology

The averages of the values for nucleotide diversity (A), the three neutrality indices (B), and Weir & Cockerham's F_{ST} (C) for each functional category are shown. Each symbol represents the average value for each population group. For nucleotide diversity (A) and Weir & Cockerham's F_{ST} values (C), solid gray lines indicate the median values of the putative enhancers of the 166 housekeeping genes calculated for all 2,504 individuals; dash-dotted gray lines represent the median values of the enhancers of the 166 housekeeping genes calculated for Africans; the gray dotted lines, loosely-dashed lines, densely-dashed, and dashed lines indicate the median values of the enhancers of the 166 housekeeping genes calculated for Europeans, East Asians, South Asians, and Americans, respectively. For the neutrality indices (B), the gray dash-dotted line represents the median values of the putative enhancers of the 166 housekeeping genes calculated for all 2,504 individuals. The number of genes belonging to each category is shown below the name of each functional category.

Table 1. The 100 sialome SNPs with the highest degree of population differentiation

Functional Categories	# of SNPs	Genes	*P*-value
Biosynthesis	0	--	0.063
Activation, Transport, Transfer	39	*ST3GAL1*, *ST3GAL4*, *ST6GAL1* *ST6GALNAC3*, *ST8SIA1*, *ST8SIA2*	1
Recognition	56	*LAMA1*, *LAMA2*, *SELP*, *SIGLEC5*	**2.815e-7****
Recycling, Degradation	5	*SIAE*	0.584



1. Biosynthesis

GNE, NAGK, NANP
NANS, RENBP

2. Activation, Transport Transfer

CMAS, SLC35A1, ST3GAL1-6
ST6GAL1-2, ST6GALNAC1-6
ST8SIA1-6

3. Modification

CMAH

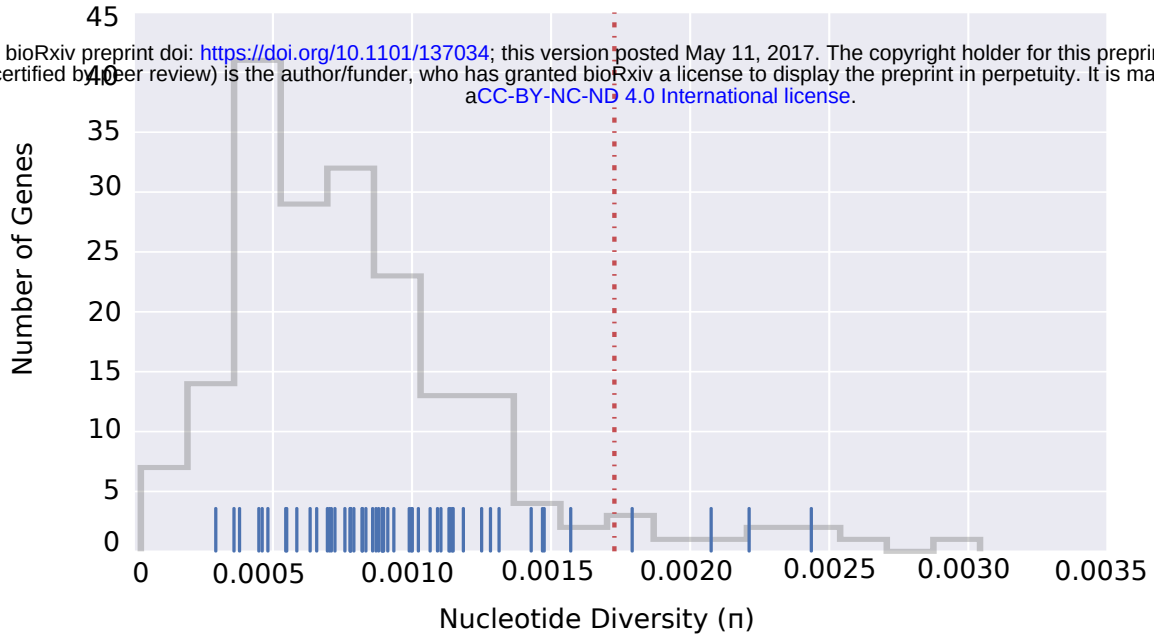
4. Recognition

SIGLEC1, CD22, CD33, MAG
SIGLEC5-12,14-16, LAMA1-2
SELE, SELL, SELP, CFH
L1CAM

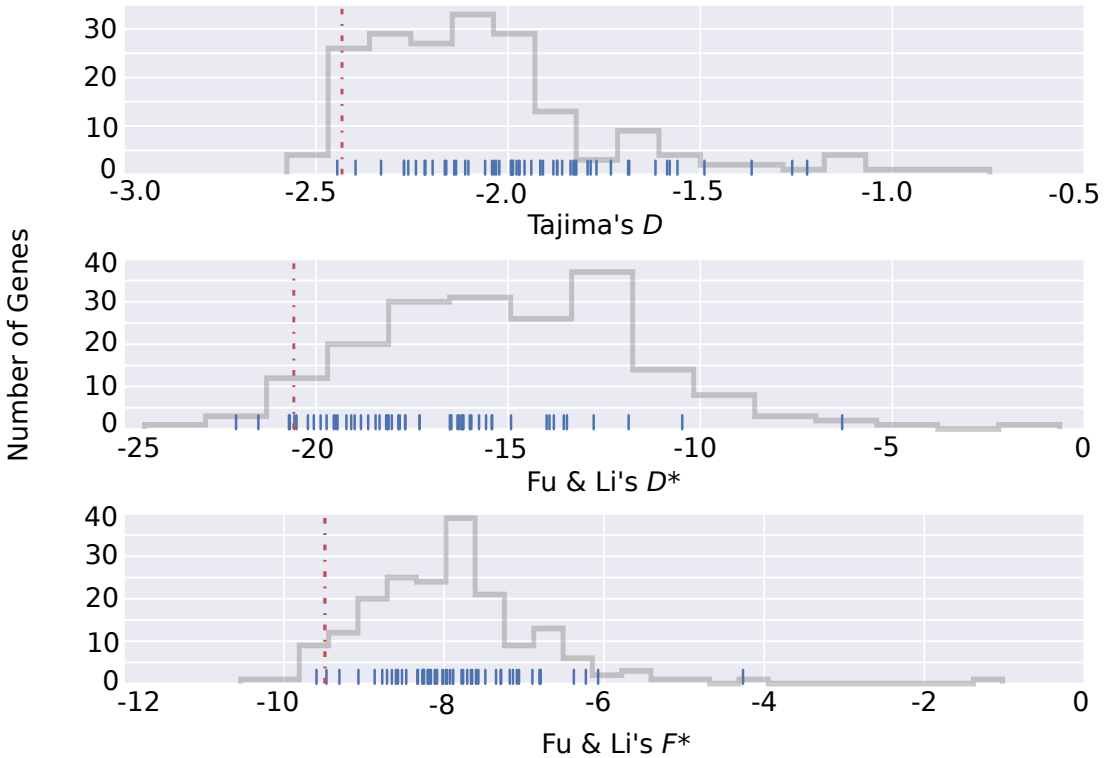
5. Recycling, Degradation

NEU1-4, SLC17A5, SIAE
NPL, CTSA

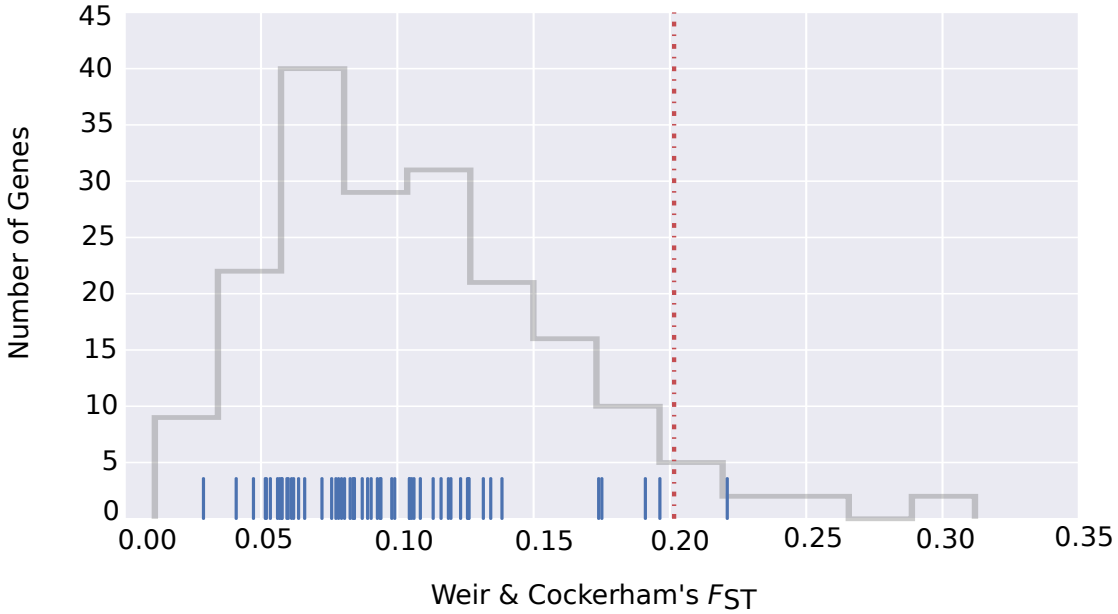
A

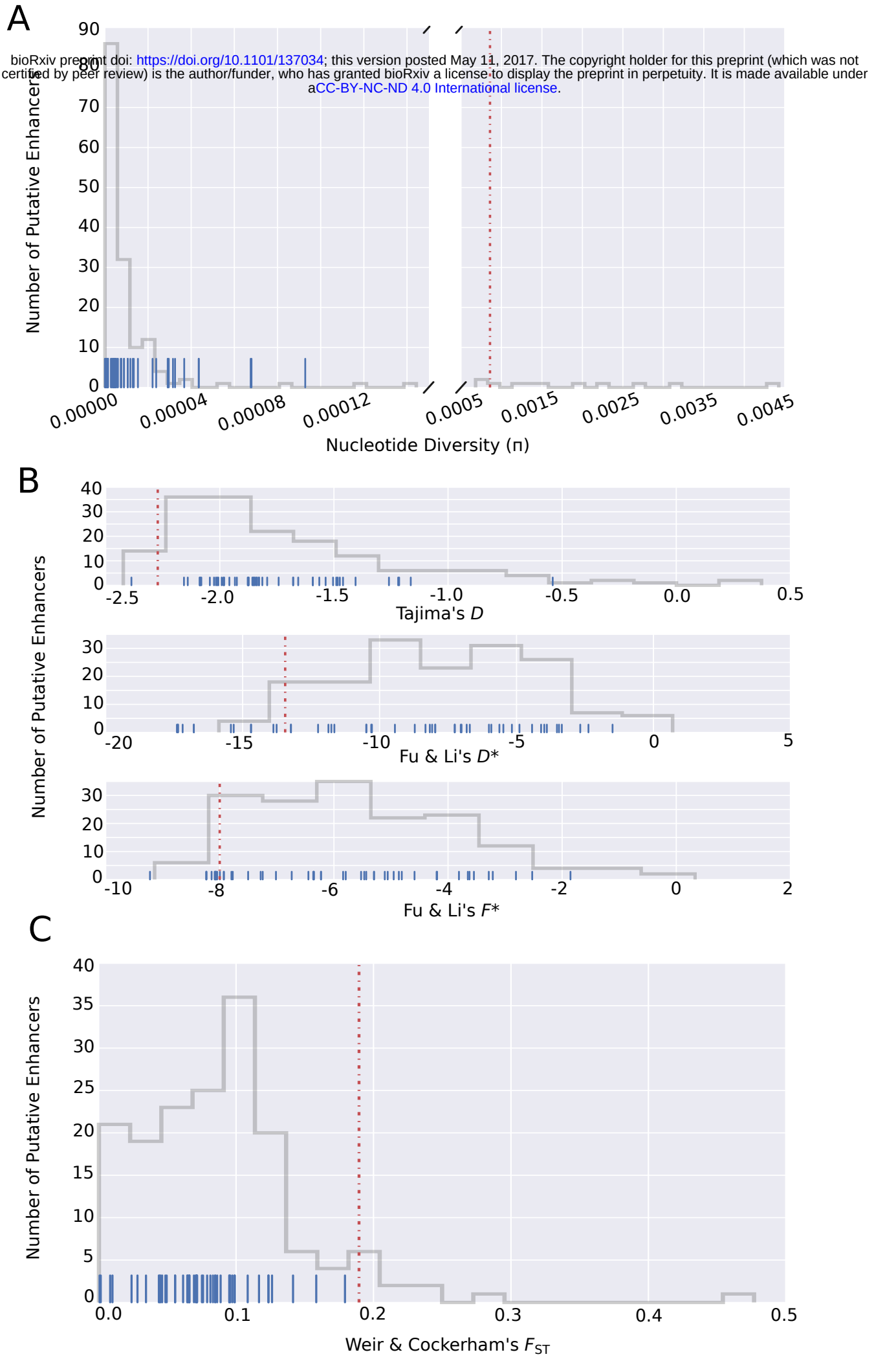


B

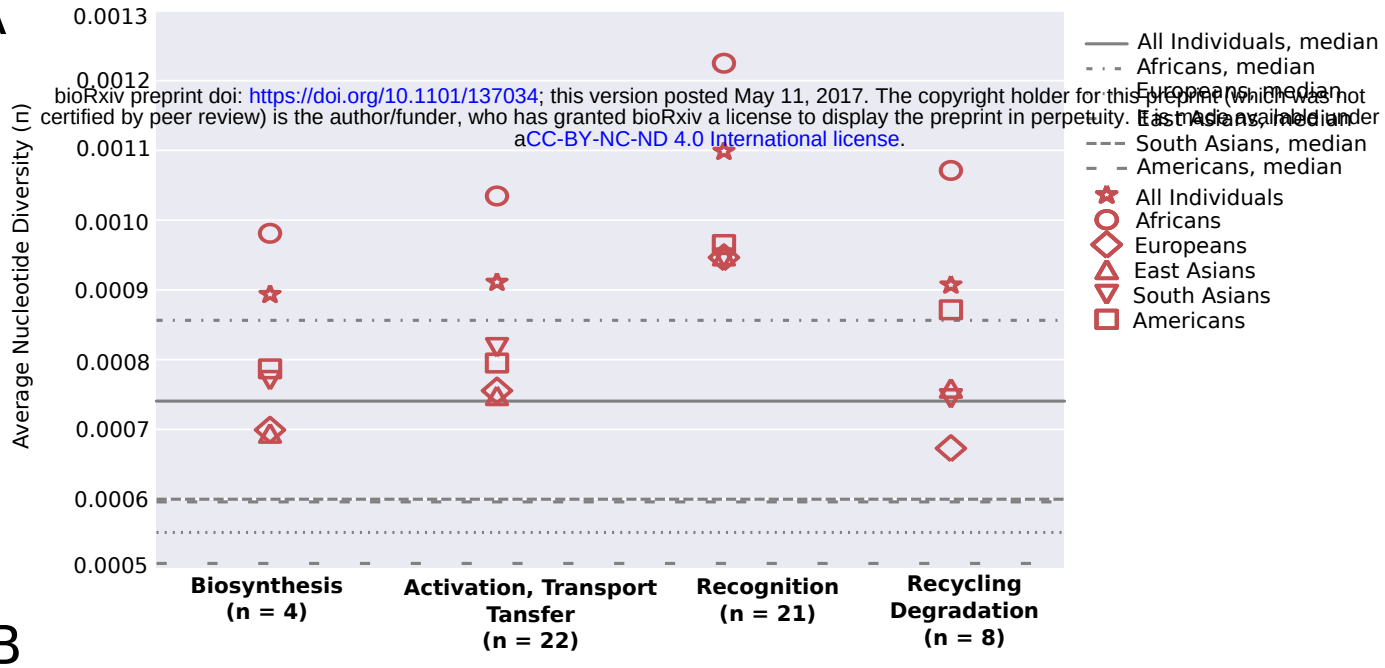


C

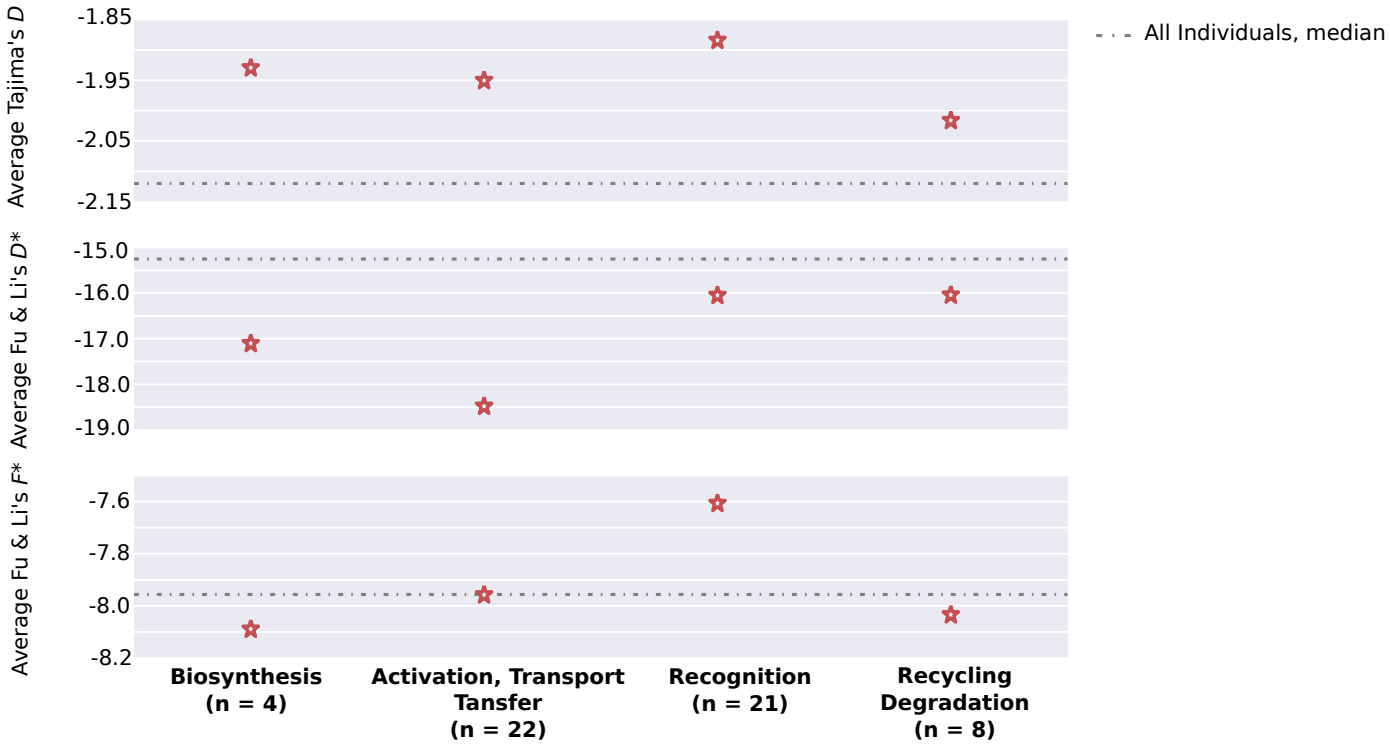




A



B



C